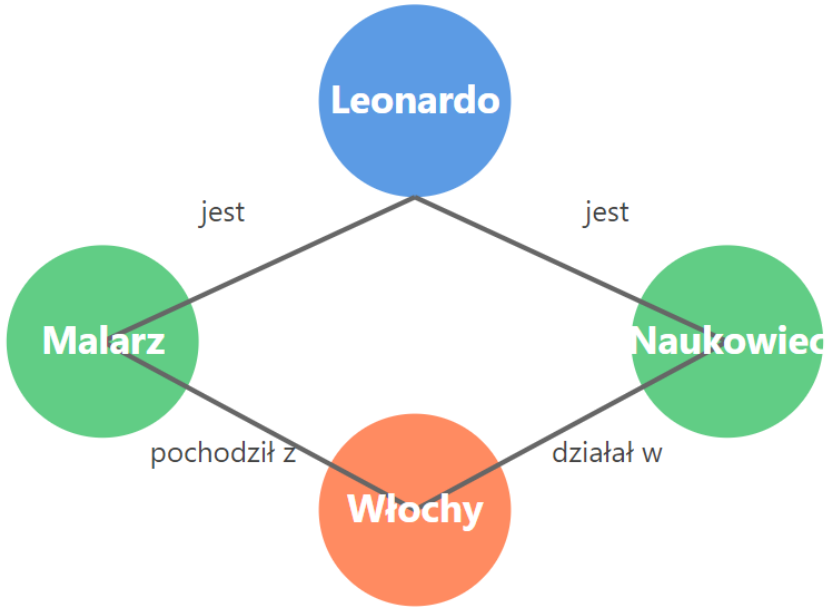
Wykorzystanie generatywnej AI i grafów wiedzy w skomplikowanych sporach sądowych

Jesienne spotkanie z legaltechem – 6.11.2024

Rafał Bałdys - Rembowski

co to są grafy wiedzy?

- Graf wiedzy to mapa informacji pokazująca powiązania między informacjami - podobnie jak sieć połączonych ze sobą punktów, gdzie każdy punkt to konkretna informacja (jak osoba, miejsce czy pojęcie), a linie między nimi pokazują, jak te informacje są ze sobą związane.

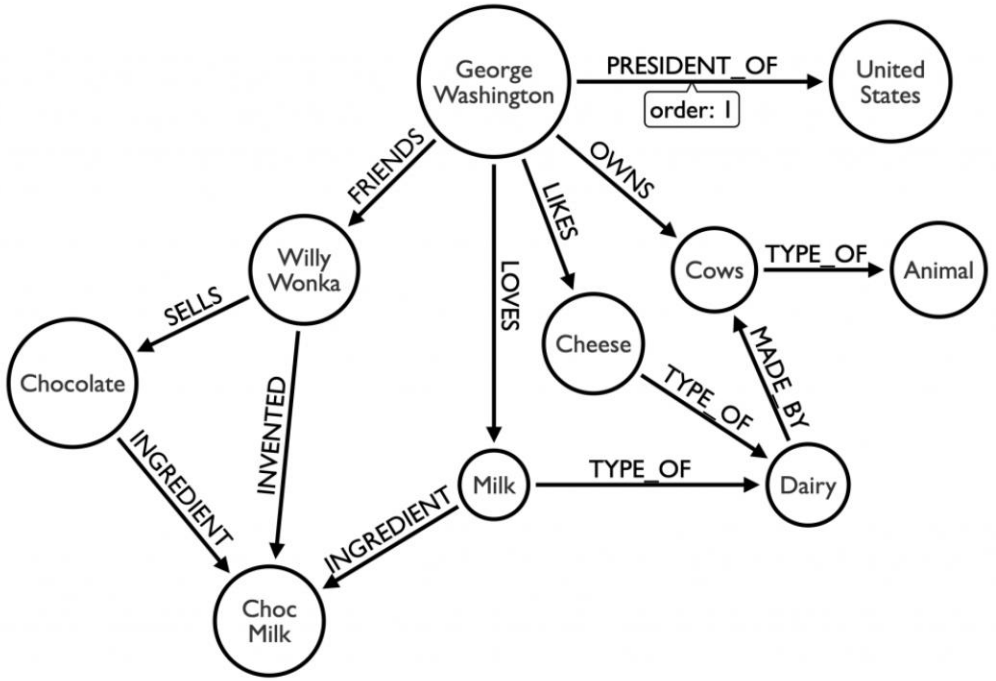


znamy to z filmów

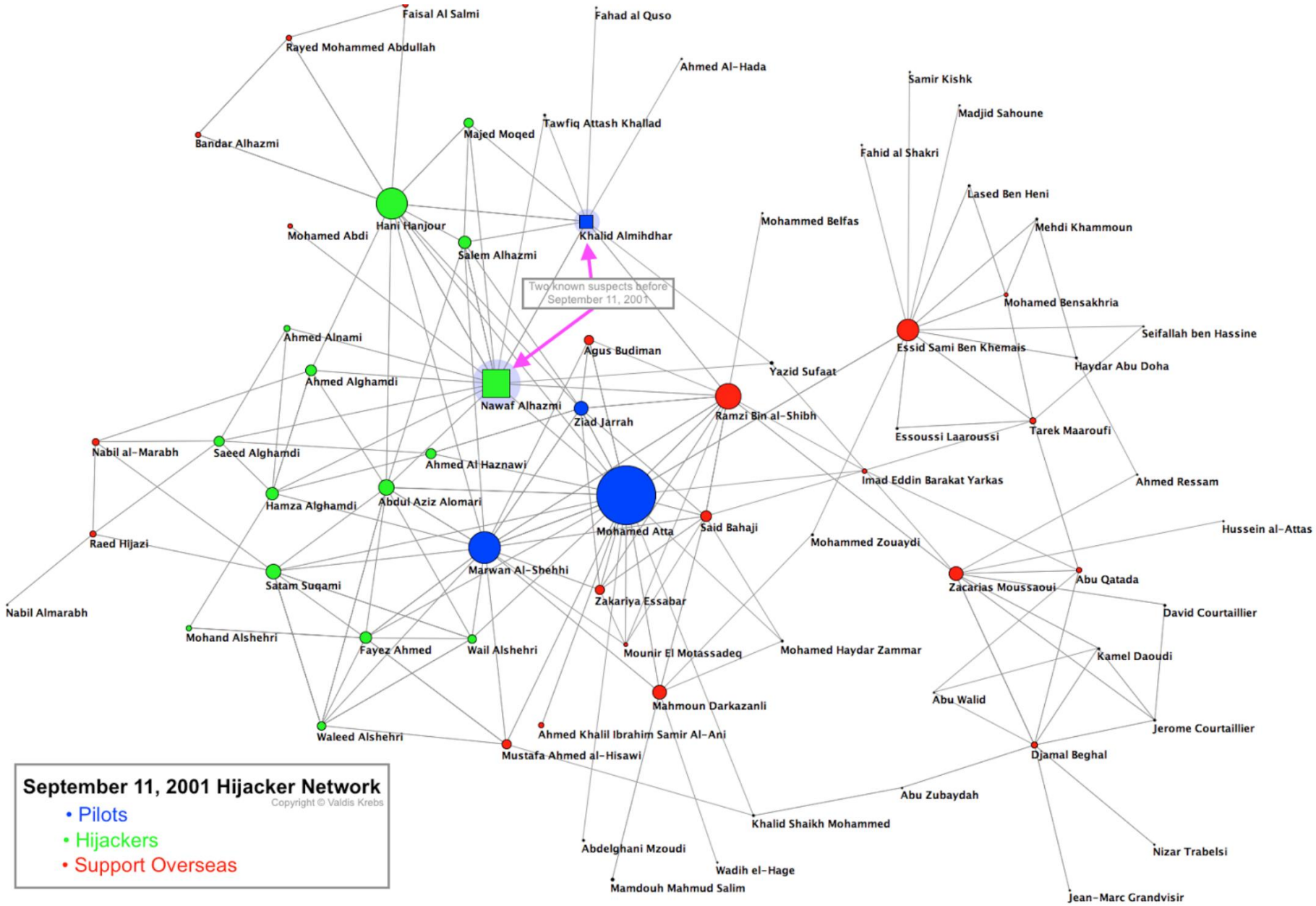
Piękny Umysł (2001) / Sherlock Holmes: Gra cieni (2012)

Kluczuwe pojęcia:

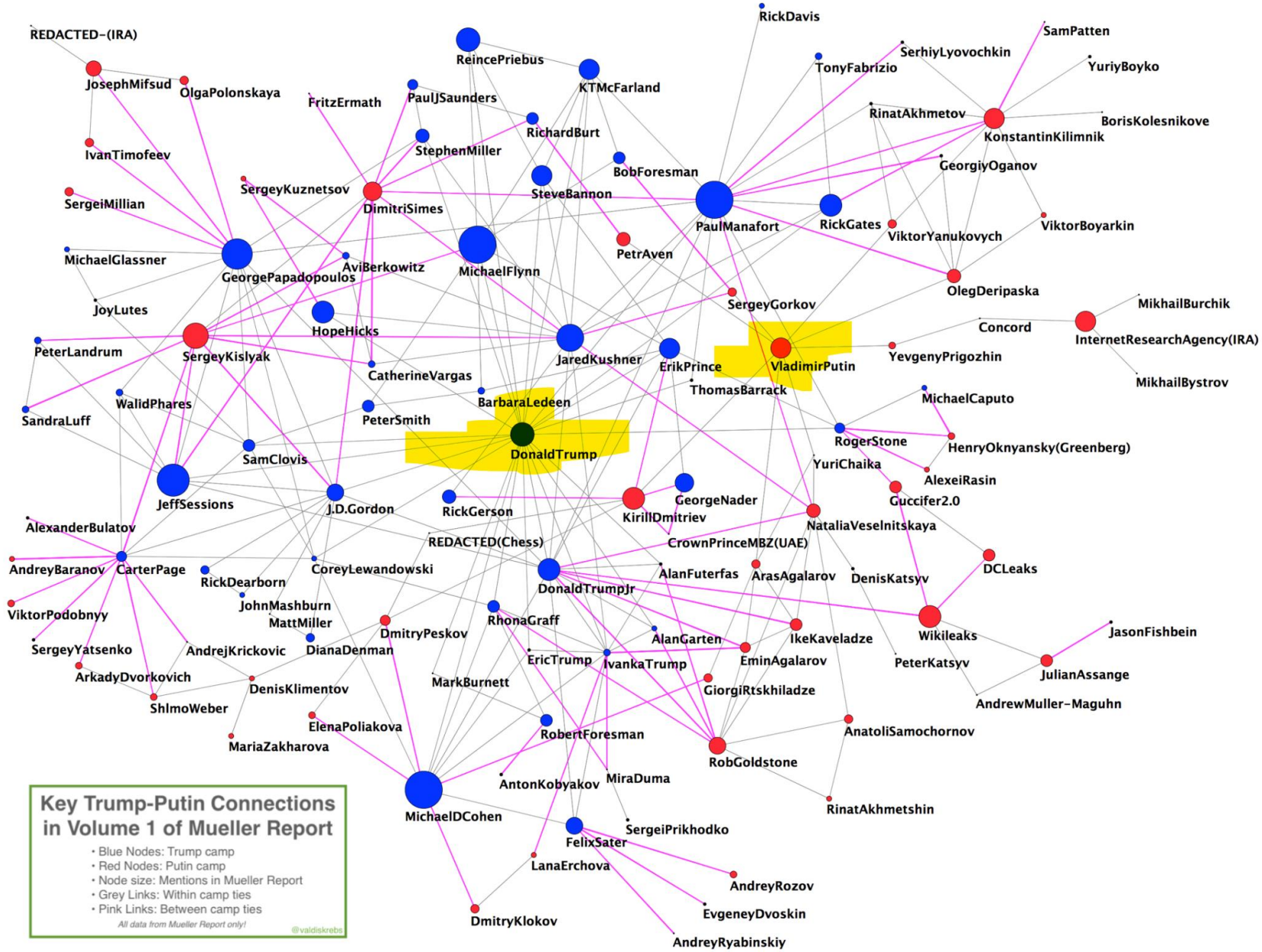
- Węzły (NODES) – osoby, dokumenty, miejsca, proces, zdarzenie
- Krawędzie (EDGES / LINKS) – relacje między nimi (należy, wynika, powoduje, etc.)



początki to lata 60te XX wieku, ale pierwsze sukcesy GW to rozpracowanie siatki porywaczy Al-Qaidy od 2012 roku pojawia się w algorytmach Google



Corruption Networks



jak robi się grafy?

i tu wchodzą LLM’y (całe na biało)

- Modele językowe nie męczą się przy czytaniu treści (choć kosztują tokeny)
- Nawet małe (tanie i szybkie) modele językowe potrafią skutecznie „wyciągać” z treści tzw. encje, czyli *etntities* (osoby, adresy, rzeczy, dokumenty, etc.) oraz relacje między nimi.
- Definiujemy, co nas interesuje i model z pomocą odpowiednich narzędzi wyciąga z treści wszystkie relacje.

….eee, ale w zasadzie to po co?

- Jeżeli chcemy, żeby AI (modele językowe) pracowała na naszych danych musimy jakoś obejść ograniczenia technologii. Główne ograniczenie to tzw. context window, oraz lost in the middle problem.
- Okno kontekstowe to maksymalna liczba słów/tokenów, jaką model językowy może "zobaczyć" i przetworzyć w jednym momencie - to jakby "pamięć robocza" modelu. Na przykład, jeśli model ma okno kontekstowe 4000 tokenów, może przeanalizować tekst mniej więcej długości 3000 słów na raz. (każde pytanie-odpowiedź musi się zmieścić w oknie kontekstowym)
- Modele językowe „tracą uwagę” przy długich treściach (podobnie jak ludzie)
- Grafy wiedzy pozwalają na zbudowanie złożonej „pamięci” wszystkich ważnych encji w danym kontekście (zapobiega pogubieniu i halucynacjom)

AI bez RAG jest „zwykłym” Chatem

Są trzy metody (obecnie) na poplątanie AI z własnymi danymi:

- RAG (Retrival Augmented Generation) – sprawdza się dla 60% use-casów
- GraphRAG (potocznie), sprawdza się tam, gdzie nie sprawdza się klasyczna metoda.
- Hybrid RAG (połączenie obu metod)
- NEW! W zeszłym tyg. IBM udostępnił nowe modele GRANITE (open-source) i nową technologię zasilania modelu własną wiedzą (jeszcze tego nie testowałem)

Nie tylko generowanie treści – wartość niesie sam graf wiedzy.

- Zbudowałem graf wiedzy dla odpowiedzi na pozew inwestora.
- Graf pozwolił mi na ocenę tego, jak złożona jest sprawa. Wiele osób (w tym ja) ma problem z oceną ilości pracy koniecznej do poświęcenia sprawie.
- Dokument wejściowy ma 116 stron (duży spór)

Co widać, czego nie widać

Widać:

- Złożoność – na 1 rzut oka widać
- Możliwość „ustawienia widoku”, żeby uchwycić interesujące nas relacje

Nie widać:

- Spójność i powtarzalność odpowiedzi – ten setup praktycznie nie halucynuje
- Możliwość generowania odpowiedzi w odniesieniu do ilości danych nieporównywalnych do okna kontekstowego.
- Możliwość jednoczesnego łączenie różnych źródeł wiedzy (np. repozytorium firmy / kancelarii)

Prezentacja grafu dla case’u (repo w komentarzu)